

Statistical Inference for Dynamical Systems: A review

Kevin McGoff, Sayan Mukherjee, Natesh S. Pillai

Abstract. The topic of statistical inference for dynamical systems has been studied extensively across several fields. In this survey we focus on the problem of parameter estimation for nonlinear dynamical systems. Our objective is to place results across distinct disciplines in a common setting and highlight opportunities for further research.

1. INTRODUCTION

The problem of parameter estimation in dynamical systems appears in many areas of science and engineering. Often the form of the model can be derived from some knowledge about the process under investigation, but parameters of the model must be inferred from empirical observations in the form of time series data. As this problem has appeared in many different contexts, partial solutions to this problem have been proposed in a wide variety of disciplines, including nonlinear dynamics in physics, control theory in engineering, state space modeling in statistics and econometrics, and ergodic theory and dynamical systems in mathematics. One purpose of this study is to present these various approaches in a common language, with the hope of unifying some ideas and pointing towards interesting avenues for further study.

By a dynamical system we mean a stochastic process of the form $(X_n, Y_n)_n$, where X_{n+1} depends only on X_n and possibly some noise, and Y_n depends only on X_n and possibly some noise. We think of X_n as the true state of the system at time n and Y_n as our observation of the system at time n . The case when no noise is present has been most often considered by mathematicians in the field of dynamical systems and ergodic theory. In this case, all uncertainty in the system comes from the uncertainty in the initial state of the system, and the ability to estimate any parameters in the system may depend strongly on properties of the observation function $f(X_n) = Y_n$, although such questions have rarely been addressed rigorously. State space models, considered most often by statisticians, lie at the other end of the noise spectrum, where both X_{n+1} and Y_n depend on some noise. Hidden Markov models, which have received considerable attention, provide a broad class of examples of these systems. In this setting, the statistical question of consistency for methods of parameter estimation has been studied, and some general results are available. The other two possible

Department of Mathematics, Duke University, Durham, NC USA, 27708 (e-mail: mcgoff@math.duke.edu) Departments of Statistical Science, Computer Science, and Mathematics, Duke University, Durham, NC USA, 27708 (e-mail: sayan@stat.duke.edu) Department of Statistics, Harvard University, 1 Oxford Street, Cambridge MA, USA, 02138 (e-mail: pillai@fas.harvard.edu)

assumptions on the presence of noise (assuming only dynamical noise or assuming only observational noise) have received relatively little attention, especially from the statistical point of view. While many of the proposed methods of parameter estimation for dynamical systems with observational noise have been studied via numerical simulations or on particular data sets, very few of these methods have been studied on a theoretical level. In fact, in the observational noise setting, basic statistical questions, such as whether a proposed method is consistent, have been considered only rarely, if at all, despite the fact that such systems capture important features of many experimental settings.

Consider, for example, the question of parameter inference for models of gene regulatory networks. The underlying model often favored by biologists consists of a system of ordinary differential equations, with each variable in the state vector representing the expression level of a particular gene in the network. For some networks of interest, a significant amount of work has produced biological understanding regarding the qualitative interactions between the genes in the network, but the corresponding ODE models still contain several parameters necessary for quantifying these interactions. Experimentalists are able to conduct experiments in which the expression levels of the genes in the network are measured at regularly spaced instances of time. The resulting data may be interpreted as time series data generated by a system of ODEs with noisy observations. The parameter inference problem in this setting consists of inferring the parameters of the ODE model from the observed data, and to the best of our knowledge there are no general statistical inference schemes for this type of problem that have been shown to be consistent.

Another example of interest is identifying the behavior of a dynamical system on a network. In a variety of applications one considers nodes in a communication network and measures the states of these nodes (or properties of the nodes) over time. In many settings, one would like to detect drastic changes in the nature of the dynamic behavior of the system. This problem is of vital importance to a variety of security applications on networks, and it can be formalized as the inference of large changes in the parameters of the network – a change point model for a dynamic network.

The objective of this article is to survey methodology across a variety of fields for parameter inference in stochastic dynamical systems. We first state the various goals of inference in dynamical systems. Our focus will be parameter inference and we provide a natural decomposition of parameter inference into four possible settings defined by the structure of noise in the system. We then state what is known in terms of rigorous results for parameter inference in these four settings. Of these settings the case of deterministic dynamics with observational noise is the least developed in terms of sound statistical theory and will be our focus. We also mention several important open problems for parameter inference in these types of systems.

There is an extremely large body of work stretching across many disciplines that relates to the topic of statistical properties of dynamical systems. Although we attempt to provide references when possible, we make no attempt to be exhaustive, and we recognize that in fact many references have been omitted. On the other hand, we hope that the references cited in this article may serve as an appropriate starting point for further reading.

2. BASIC DEFINITIONS AND PRELIMINARIES

The most general setting that we will consider may be described as follows. Let \mathcal{A} , \mathcal{X} , \mathcal{Y} , and \mathcal{N} be Polish¹ spaces (complete metric spaces with a countable dense set), where each one is equipped with its Borel σ -algebra. The space \mathcal{A} denotes the parameter space, the underlying dynamical system evolves in the space \mathcal{X} and the observations take values in \mathcal{Y} . We consider a stochastic process $(X_n, Y_n)_n$, which satisfies the following dynamics: for some a in \mathcal{A} and X_0 distributed according to a Borel probability measure on \mathcal{X} ,

$$(2.1) \quad Y_n = f_a(X_n, \epsilon_n),$$

$$(2.2) \quad X_{n+1} = T_a(X_n, \delta_{n+1}),$$

where δ_{n+1} is the dynamical noise and ϵ_n is the observational noise. The maps $T : \mathcal{A} \times \mathcal{X} \times \mathcal{N} \rightarrow \mathcal{X}$ and $f : \mathcal{A} \times \mathcal{X} \times \mathcal{N} \rightarrow \mathcal{Y}$ determine the evolution of the state space dynamics and the observation process, respectively. We refer to a sequence $(X_n)_n$ satisfying (2.2) as a trajectory and a sequence $(Y_n)_n$ satisfying (2.1) as a sequence of observations.

Let us start with the following definitions.

DEFINITION 2.1. A stochastic process $(X_n)_n$ is stationary if for any k , n and n_1, \dots, n_k in \mathbb{N} , the joint distribution of $(X_{n_1+n}, \dots, X_{n_k+n})$ is equal to the joint distribution of $(X_{n_1}, \dots, X_{n_k})$.

DEFINITION 2.2. An \mathcal{X} -valued stationary stochastic process $(X_n)_n$ is said to be ergodic if for every $\ell \geq 1$ and every pair of Borel sets $A, B \in \mathcal{X}^\ell$,

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n \mathbb{P}\left((X_1, \dots, X_\ell) \in A, (X_{k+1}, \dots, X_{k+\ell}) \in B\right) \\ = \mathbb{P}\left((X_1, \dots, X_\ell) \in A\right) \mathbb{P}\left((X_1, \dots, X_\ell) \in B\right). \end{aligned}$$

DEFINITION 2.3. A measurable dynamical system is a triple $(\mathcal{X}, \mathcal{F}, T)$, where $(\mathcal{X}, \mathcal{F})$ is a measurable space and $T : \mathcal{X} \rightarrow \mathcal{X}$ is measurable. A topological dynamical system is a pair (\mathcal{X}, T) , where \mathcal{X} is a topological space and $T : \mathcal{X} \rightarrow \mathcal{X}$ is a continuous map. In the study of topological dynamics, one often assumes that \mathcal{X} is compact and metrizable.

DEFINITION 2.4. A measure-preserving system is a quadruple $(\mathcal{X}, \mathcal{F}, T, \mu)$, where $(\mathcal{X}, \mathcal{F}, \mu)$ is a measure space, $T : \mathcal{X} \rightarrow \mathcal{X}$ is measurable, and $\mu(T^{-1}(A)) = \mu(A)$ for each A in \mathcal{F} . In this case, we say that T preserves the measure μ and μ is an invariant measure for T . For the purpose of this article, we will always assume that any invariant measure μ is a probability measure. Also, if \mathcal{X} is Polish and \mathcal{F} is the Borel σ -algebra, then we may refer to (\mathcal{X}, T, μ) as a measure-preserving system.

DEFINITION 2.5. A measure-preserving system $(\mathcal{X}, \mathcal{F}, T, \mu)$ is ergodic if $T^{-1}(A) = A$ implies $\mu(A) \in \{0, 1\}$ for any A in \mathcal{F} . We may say that T is ergodic for μ , or we may say that μ is ergodic for T .

¹This is a classical assumption in dynamical systems.

With the definitions given above, there is a correspondence between stationary stochastic processes and measure-preserving systems. Let us describe this correspondence as follows. Suppose $(X_n)_n$ is an \mathcal{X} -valued stationary stochastic sequence, where \mathcal{X} is Polish. Let $\mathcal{Y} = \prod_n \mathcal{X}$, equipped with the product σ -algebra induced by the Borel σ -algebra on \mathcal{X} . Define $T : \mathcal{Y} \rightarrow \mathcal{Y}$ by the left shift: if $y = (x_n)_n$, then $(T(y))_n = x_{n+1}$. Kolmogorov's consistency theorem gives that there is a unique probability measure μ on \mathcal{Y} with the same finite dimensional distributions as $(X_n)_n$. In this case, the stationarity of $(X_n)_n$ corresponds exactly to the invariance of μ with respect to T . Moreover, if $(X_n)_n$ is ergodic, then μ is ergodic for T .

In the other direction, given any measure-preserving system (\mathcal{X}, T, μ) , we may define a stationary stochastic process as follows. For any Polish space \mathcal{Y} and measurable map $f : \mathcal{X} \rightarrow \mathcal{Y}$, let $X_n(\omega) = f(T^n(\omega))$. If (\mathcal{X}, T, μ) is ergodic, then so is $(X_n)_n$.

Recall that an \mathcal{X} -valued stochastic process $(X_n)_n$ is a Markov chain if for every x in \mathcal{X} , there exists a probability measure $\pi(x, \cdot)$ on \mathcal{X} such that for each measurable set A in \mathcal{X} , it holds that

$$\mathbb{P}(X_{n+1} \in A | X_1 = x_1, \dots, X_n = x_n) = \pi(x_n, A).$$

In the model (2.1)-(2.2), if the dynamical noise process $(\delta_n)_n$ is assumed to be i.i.d., then both $(X_n)_n$ and (X_n, Y_n) are Markov chains. This fact is particularly relevant in Sections 5 and 6, where the process $(\delta_n)_n$ is assumed to be non-zero. Even in this case the process $(Y_n)_n$ may exhibit long-range dependencies. Setting the dynamical noise to zero in model (2.1)-(2.2) can be thought of as a very degenerate Markov chain, but it is not clear in this case how helpful the Markov perspective is, since even the process $(X_n)_n$ may exhibit long-range dependencies.

2.1 Goals of statistical inference

There are a variety of topics that can be considered part of “statistical inference in dynamical systems.” In the interest of providing context for this survey, let us mention the following topics:

1. parameter estimation, model identification or reconstruction;
2. state estimation, filtering, smoothing, or denoising;
3. feature estimation, where features often include invariant measures, dimensions, entropy, or Lyapunov exponents;
4. prediction or forecasting;
5. noise quantification, estimation, or detection.

In this paper we focus almost exclusively on the problems of parameter inference, system identification or reconstruction. In the setting of (2.1)-(2.2), we pose the parameter estimation problem as follows. Suppose the family of dynamical systems can be parametrized by T_a , with parameter $a \in \mathcal{A}$, as in (2.2). Construct statistical procedures for estimating the parameter a , given observations Y_1, Y_2, \dots, Y_n from (2.1), and provide adequate theoretical support for the validity of the estimation procedure.

Of course, the boundaries between the problems mentioned above are often quite blurred. For example, if one can accurately estimate the hidden states

$(X_k)_{k=0}^{n-1}$ from the data $(Y_k)_{k=0}^{n-1}$, then the problem of system identification often becomes significantly easier. For this reason, parameter inference methods often simultaneously attempt some version of state estimation or denoising.

2.2 Organization of the paper

We organize this survey according to which of the two types of noise in (2.1)-(2.2) are present (*i.e.* non-zero). This organization is motivated by the observation that methods and results for parameter inference in dynamical systems tend to be specific to the type of noise assumed in the model.

The remainder of the paper is organized as follows. In Section 3 we describe some results relevant to inference for dynamical systems in the absence of noise. Section 4 contains a variety of proposed methods dealing with the case of dynamical systems contaminated by observational noise only. Section 5 deals with the case of only dynamical noise, and Section 6 addresses the setting of state space models, that is systems with both dynamical and observational noise. Lastly, we highlight some possibly interesting open questions in Section 7.

Ornstein and Weiss [91] have shown that in a certain sense it is impossible, in general, to tell the difference between observational and dynamical noise. In this sense, one might suggest that from the point of view of abstract ergodic theory, we should not make distinctions on the basis of the type of noise present. However, we are interested in finer properties than those captured by the equivalence relations considered in [91], and therefore the distinction between observational and dynamical noise might still be useful for our purposes.

2.3 Related surveys and books

There have been many other reviews of topics related to the topics in this survey. An incomplete list of such reviews is the following: [7, 9, 15, 30, 45, 50, 66, 114, 120]. Furthermore, let us mention the following books or monographs related to the topics in this survey: [1, 8, 14, 29, 62, 64, 91, 118]. The relevance of this survey is that we bring together approaches from many distinct fields and discuss them in a common statistical setting. In particular we discuss parameter estimation and inference for the full range of noise settings. This perspective is rare since the different noise settings often correspond to different research areas such as deterministic dynamics or state space methods based on hidden Markov models. We bring these various approaches together and place them in a common context. Inference in dynamical systems for a variety of contexts was discussed in Berliner (1992) [7], and our survey can be thought of as an updated and greatly expanded version of this work.

3. NO NOISE

If no noise is present in the model (2.1)-(2.2), then we have the following situation:

$$(3.1) \quad Y_n = f_a(X_n)$$

$$(3.2) \quad X_{n+1} = T_a(X_n),$$

where $T : \mathcal{A} \times \mathcal{X} \rightarrow \mathcal{X}$ is a parametrized family of maps and $f : \mathcal{A} \times \mathcal{X} \rightarrow \mathcal{Y}$ is a parametrized family of observation functions. For a fixed parameter value a , the model (3.1)-(3.2) is one of the classical objects of study in dynamical systems and

ergodic theory (for general references on dynamical systems and ergodic theory, see [10, 63, 97, 123]).

3.1 Non-parametric system reconstruction from direct observations

Here we consider non-parametric estimation of a map T from direct observation of a single trajectory. Although the methods discussed in this section do not directly involve parameter estimation, they are nonetheless relevant for parameter estimation, since any non-parametric method for estimation of a map immediately yields a method of parameter estimation if the map to be estimated comes from a parameterized family.

Let us first consider a case when the system can be successfully reconstructed from observations. If \mathcal{X} is a manifold, T is continuous, the trajectory $(x_n)_n$ is dense in \mathcal{X} , and we observe the trajectory directly (*i.e.* the observations $(y_n)_n$ satisfy $x_n = y_n$), then T can be consistently estimated from $(y_n)_n$ using locally linear functions of the data. More precisely, let us state a result from [4] justifying this statement in the case $\mathcal{X} = [0, 1]$. Let λ be Lebesgue measure on $[0, 1]$. The map $T : [0, 1] \rightarrow [0, 1]$ is said to be an $E\{I_j, \alpha_j\}$ -map if there exists at most countably many disjoint open intervals I_j and real numbers α_j such that $\lambda(\cup I_j) = 1$ and $f'(x) = \alpha_j$ for all x in I_j .

PROPOSITION 3.1 ([4]). *Let T be an $E\{I_j, \alpha_j\}$ -map. Suppose the observed trajectory $(x_n)_n$ is dense in $[0, 1]$. Then there exists a sequence of estimates \hat{T}_n of T such that for almost every x in $[0, 1]$, it holds that $\hat{T}_n(x) = T(x)$ for all but finitely many n . In particular, \hat{T}_n converges to T pointwise almost everywhere, and $\lambda(\{x : \hat{T}_n \neq T(x)\})$ tends to zero.*

To get an idea about how to prove this proposition, notice that for any two consecutive points x_n and x_{n+1} in the trajectory, the pair (x_n, x_{n+1}) lies on the graph of T . Therefore one may estimate T by linearly interpolating between neighboring points on the graph of T .

When the map T is not assumed to be continuous but only measurable, estimation of T from discrete observations of a single trajectory has been carried out by Adams and Nobel [4]. In this work, the map T is assumed to preserve a Borel probability measure μ on \mathcal{X} , and the system (X, μ, T) is assumed to be ergodic. Their main result may be stated as follows.

THEOREM 3.2 ([4]). *Let μ_0 be a reference probability measure on \mathcal{X} that is assumed to be “known.” Also assume that there is a “known” constant M such that $1/M \leq d\mu/d\mu_0 \leq M$. Let $\text{Meas}(\mathcal{X})$ denote the space of measurable functions from \mathcal{X} to \mathcal{X} . Then there is an estimation scheme $(T_n)_n$ (whose definition uses M and μ_0), where $T_n : \mathcal{X}^n \rightarrow \text{Meas}(\mathcal{X})$, such that for μ_0 -a.e. initial condition x_0 , the map $T_n(x_0, \dots, x_{n-1})$ converges to T in a weak topology (*i.e.* $\mu(T_n^{-1}(A) \Delta T^{-1}(A))$ tends to zero as n tends to infinity for each Borel set A).*

The estimation scheme $(T_n)_n$ that appears in [4] is constructed using an adaptive histogram method, which we discuss below. This paper also shows that under the same hypotheses the conclusion of the theorem is false if one requires that $\mu(\{x \in \mathcal{X} : T_n(x) \neq T(x)\})$ tends to zero as n tends to infinity.

Here we give an idea of the estimation scheme used in the proof of Theorem 3.2. The histogram method described here is actually from [88], which is very similar in spirit to the method used in the proof of Theorem 3.2. Assume that $\mathcal{X} \in \mathbb{R}^d$, and we fix a refining sequence $(\pi_k)_k$ of finite partitions of \mathcal{X} with some additional properties (see [4] for details). Let $\pi_k(x)$ denote the cell in π_k containing x . Given the first n terms of the trajectory $(x_j)_{j=0}^{n-1}$, let

$$\phi_{n,k}(x) = \frac{\sum_{j=0}^{n-1} x_{j+1} I_{\{x_j \in \pi_k(x)\}}}{\sum_{j=0}^{n-1} I_{\{x_j \in \pi_k(x)\}}},$$

where $I_{\{x_j \in \pi_k(x)\}}$ is the indicator function of the event that x_j is in $\pi_k(x)$, and if the cell $\pi_k(x)$ contains no points x_j , then $\phi_{n,k}(x) = 0$. Now consider the empirical loss of $\phi_{n,k}$:

$$\Delta_{n,k} = \left(\frac{1}{n} \sum_{j=0}^{n-2} (\phi_{n,k}(x_j) - x_{j+1})^2 \right)^{1/2}.$$

The estimates \hat{T}_n of T are adaptively chosen from among the $\phi_{n,k}$ according to $\Delta_{n,k}$ (using μ_0 and M). This method has the advantage that it works in quite a general setting (the only assumptions involve ergodicity and the Radon-Nikodym derivative with respect to a reference measure). On the other hand, it relies on the ergodic theorem for convergence, and therefore it appears very unlikely that it would have any general speed of convergence.

3.2 Non-parametric system reconstruction from general observations

In this section we consider approaches to system reconstruction when the observations $(y_n)_n$ are not necessarily equal to the trajectory $(x_n)_n$. There is a vast amount of literature on the technique of system reconstruction via delay coordinate embeddings. These system reconstructions may be thought of as non-parametric inference of dynamical systems. Delay coordinate embeddings are a well-studied inference procedure to reconstruct dynamical systems that satisfy certain conditions. In this section we define delay coordinate embeddings, mention some of the main uses of these techniques, and provide some representative theorems that provide conditions under which these methods work.

The eventual goal of delay coordinate embedding techniques is typically feature estimation, which we summarize as follows. If the underlying map T and the observation function are both smooth, then under generic conditions, a delay coordinate embedding allows one to construct a smooth map \tilde{T} such that \tilde{T} is related to T by a smooth change of coordinates. Under this scenario, T and \tilde{T} will share many features, including entropy, Lyapunov exponents, and fractal dimensions of corresponding invariant measures. As these features are considered important in many physical settings, such delay coordinate reconstructions have been extensively studied.

To be specific, we consider a smooth map $T : \mathcal{X} \rightarrow \mathcal{X}$ of a manifold \mathcal{X} , with a smooth observation function $f : \mathcal{X} \rightarrow \mathbb{R}$. The data are assumed to be generated as follows: there is a trajectory $(x_n)_n$ such that $x_{n+1} = T(x_n)$, and we observe the data $(y_n)_n$ such that $y_n = f(x_n)$. The original idea to use delay coordinate embeddings to construct a system equivalent to (\mathcal{X}, T) from the observations is due to Ruelle, at least according to the influential paper [94].

DEFINITION 3.3. A delay coordinate mapping of \mathcal{X} into \mathbb{R}^m is a mapping $F : \mathcal{X} \rightarrow \mathbb{R}^m$ such that

$$F(x) = (f(x), f \circ T^\tau(x), \dots, f \circ T^{\tau(m-1)}(x)),$$

for some natural number τ . The mapping F is said to be an embedding if it is a diffeomorphism from \mathcal{X} to its image $F(\mathcal{X})$, that is if F is a smooth injection and has a smooth inverse.

The well-known theorem of Takens [116] (often called the Takens Embedding Theorem) may be stated as follows.

THEOREM 3.4 ([116]). *If T , f , and τ satisfy certain genericity conditions and $m > 2 \dim(\mathcal{X})$, then F is an embedding.*

Let $\tilde{\mathcal{X}} = F(\mathcal{X})$ and $\tilde{T} = F \circ T \circ F^{-1}$. The fact that F is an embedding means that the system (\mathcal{X}, T) is related to the system $(\tilde{\mathcal{X}}, \tilde{T})$ by a smooth change of coordinates (given by F). In particular, invariants of (\mathcal{X}, T) that depend on the differential structure of T (such as Lyapunov exponents or fractal dimensions of attractors) are equal to those of the system $(\tilde{\mathcal{X}}, \tilde{T})$.

In particular, given the data $(y_k)_{k=0}^{n-1}$, we may build time series data $(s_k)_{k=0}^{n-1-\tau(m-1)}$ for the system $(\tilde{\mathcal{X}}, \tilde{T})$ as follows: for $k = 0, \dots, n-1-\tau(m-1)$, let

$$s_k = (y_k, y_{k+\tau}, \dots, y_{k+\tau(m-1)}).$$

Then the new time series $(s_k)_k$ may be used to estimate invariant features of $(\tilde{\mathcal{X}}, \tilde{T})$, which will be the same as those features of (\mathcal{X}, Q) .

Takens's theorem has been generalized in various directions, such as filtered delay embeddings (see [109], for example) or delay embeddings for stochastic systems (see [113]), but we do not attempt to record all such results. However, the following generalization, due to Sauer, Yorke, and Casdagli, bears mentioning.

THEOREM 3.5 ([109]). *Let A be a compact subset of \mathcal{X} with box-counting dimension d . Let $m > 2d$. Suppose T , f , τ , and A satisfy certain genericity conditions. Then the delay coordinate map F given above is an injection on A and an immersion on each compact subset of any smooth manifold contained in A .*

The advantage of this theorem over the Takens theorem is that the relevant dimension d might be less than the ambient dimension of \mathcal{X} , in which case the number of coordinates m required in the embedding space may be less than the number of coordinates required by Takens's theorem.

In order to use the delay coordinate method given only the data $(y_k)_{k=0}^{n-1}$, one must choose an appropriate dimension m and an appropriate lag τ . A variety of statistical techniques have been proposed to estimate the dimension m and find a suitable lag τ (for example, see the book [62] or the collection [82]), but further pursuit of these topics lies outside the scope of this survey.

3.3 Results from ergodic theory

In this section, we state some results from ergodic theory that are relevant for parameter inference.

One of the most general results in this area is due to Ornstein and Weiss [93]. In this work, the authors consider the problem of estimation of stationary ergodic processes. (Note that in the setting of (3.1)-(3.2), if X_0 is distributed according to an ergodic invariant measure for T_a , then the observation process $(Y_n)_n$ satisfies exactly these conditions.) To make this problem precise, they consider the \bar{d} metric on the space of such processes. Their main results may be stated as follows. First, they construct a procedure which, given a realization $(X_k)_{k=0}^{n-1}$ of a process $(X_k)_k$ constructs a process $Z^n = (Z_k^n)_k$. Then they show that the sequence of processes $(Z^n)_n$ converges to $(X_k)_k$ in the \bar{d} metric if and only if $(X_k)_k$ is Bernoulli. Thus, they have shown that there is a consistent estimation procedure for the class of Bernoulli processes. Furthermore, they show that no estimation procedure can be consistent for the class of all stationary ergodic processes.

In another direction, Ornstein and Weiss [92] show that entropy is the only finitely observable invariant in the following sense. Let J be a function from the class of finite-valued stationary ergodic processes to a complete separable metric space such that J is constant on isomorphism classes. The main result of [92] states that if J is finitely observable, then it must be a continuous function of the entropy. This result shows that there are strong restrictions on the possibilities for inference of isomorphism invariants.

Gutman and Hochman [41] extend the results in [92] in several ways. They give several rich families of classes \mathcal{C} of stationary ergodic processes such that if J is a finitely observable invariant on \mathcal{C} , then J is constant. They also show that for every finitely observable invariant J on the class of irrational circle rotations, J is constant on the processes arising from a full measure set of angles. In particular, there is no finitely observable invariant for irrational rotations which is complete.

There is a large body of work, often categorized as smooth ergodic theory, that seeks to understand the statistical properties of smooth (or piecewise smooth) dynamical systems. The typical setting is that one has a compact Riemannian manifold M and a smooth self-map $f : M \rightarrow M$. The manifold typically has a distinguished probability measure λ , which one may think of as volume measure on the manifold. The goal is to understand the asymptotic behavior of the trajectory $\{f^n(x)\}_n$ for λ -a.e. x . For a wide class of such systems [126], often called (non-uniformly) hyperbolic systems, there is an invariant measure μ on M such that for x in a set of positive λ -measure, the trajectories in x equidistribute with respect to μ . In such cases, the measure μ is said to be a *physical* measure. Often the measure μ has some additional properties (it has no zero Lyapunov exponents and absolutely continuous conditional measures with respect to λ on unstable manifolds), and in this case μ may be called an SRB (Sinai-Ruelle-Bowen) measure [127]. The ergodic theory of SRB measures is fairly well-studied, and many of their statistical properties have been analyzed.

A related topic that has seen a great deal of attention recently is concentration inequalities for dynamical systems [16, 17, 18, 19, 20]. These inequalities are used to study the fluctuations of observables for dynamical systems and have been shown to hold for sufficiently regular observables and a wide class of non-uniformly hyperbolic dynamical systems. Using these inequalities, it is possible

to perform statistical estimation of various features of the dynamical system. See the survey [16] for more details and precise statements.

3.4 Parameter inference via synchronization and control

Synchronization-based approaches to parameter estimation have appeared quite often in the physics and control systems literature [3, 79, 95, 102, 128]. In situations when these methods are used, it is common that no particular noise model is assumed. Indeed synchronization-based approaches are typically described as parameter inference methods in the noiseless setting, although they may be applied in other settings. The main idea of synchronization-based methods is to insert a “control” term in the defining equations of the system that allows one to incorporate the data. The parameter estimation may then be framed as a large optimization procedure in which one tries to find trajectories of the system which are close to the data.

The topic of parameter estimation in a noiseless setting is discussed directly in the work of Abarbanel, Creveling, Farsian, and Kostuk [2], and we review their approach in this section. The main issue in this context is that one only has access to the observations $(Y_n)_n$, which might “hide” some information about the underlying system. The approach taken in [2] involves synchronization of the observations and the output of a model over the relevant time window. This approach may be summarized as follows.

Suppose that \mathcal{X} is in \mathbb{R}^d and the system (3.1)-(3.2) has the following form:

$$\begin{aligned} Y_n &= X_{n,1} \\ X_{n+1,i} &= T_{a,i}(X_n), \end{aligned}$$

where $X_{n,i}$ denotes the i -th coordinate of X_n . The synchronization approach taken in [2] is to add a “control” term of the form $k(Y_n - X_{n,1})$ to first coordinate of the model as follows:

$$\begin{aligned} \tilde{X}_{n+1,1} &= T_{a,1}(\tilde{X}_n) + k(Y_n - \tilde{X}_{n,1}) \\ \tilde{X}_{n+1,i} &= T_{a,i}(\tilde{X}_n), \quad i > 1. \end{aligned}$$

For $k > 0$ large enough, the data Y_n and the first coordinate $\tilde{X}_{n,1}$ of the model trajectory will “synchronize.” With a fixed k , the authors propose to estimate the parameter a and the initial state X_0 by minimizing the following function:

$$C(a, X_0) = \sum_{j=0}^{n-1} (Y_j - \tilde{X}_{j,1})^2,$$

where the trajectory \tilde{X}_n is computed starting at $\tilde{X}_0 = X_0$. The purpose of adding the control term is to regularize the function C so that its minimum may be found efficiently. Of course, the trajectory \tilde{X}_n associated with this minimum is not a true trajectory of the original system. Therefore the authors propose a synchronization method that allows the parameter k to depend on time. In other words, they propose to minimize the cost function

$$C(a, X_0) = \sum_{j=0}^{n-1} (Y_j - \tilde{X}_{j,1})^2 + k_j^2,$$

subject to the constraints

$$\begin{aligned}\tilde{X}_{j+1,1} &= T_{a,1}(\tilde{X}_j) + k_j(Y_j - \tilde{X}_{j,1}) \\ \tilde{X}_{j+1,i} &= T_{a,i}(\tilde{X}_j), \quad i > 1.\end{aligned}$$

Although this method has been observed to work sufficiently well in practice [2], we remark that to the best of our knowledge there are no theoretical guarantees regarding the consistency or performance of this method.

4. OBSERVATIONAL NOISE ONLY

If only observational noise is present in the model (2.1)-(2.2), then the system (2.1)-(2.2) reduces to the following situation:

$$(4.1) \quad Y_n = f_a(X_n, \epsilon_n)$$

$$(4.2) \quad X_{n+1} = T_a(X_n),$$

where $(\epsilon_n)_n$ is a noise process, $T : \mathcal{A} \times \mathcal{X} \rightarrow \mathcal{X}$ is a parametrized family of maps, and $f : \mathcal{A} \times \mathcal{X} \times \mathcal{N} \rightarrow \mathcal{Y}$ is a parametrized family of noisy observation functions. Multiple authors explicitly argue for consideration of the observational noise model. For example, Judd [51] states that “the reality is that many physical systems are indistinguishable from deterministic systems, there is no apparent small dynamic noise, and what is often attributed as such is in fact model error.” Furthermore, Lalley and Nobel [71] remark that “estimation in the observational noise model has not been broadly addressed by statisticians, though the model captures important features of many experimental situations.”

A distinguishing feature of the observational noise model is that the process $(X_n)_n$ is deterministic, and therefore in general it exhibits a long-range dependence structure. Furthermore, this long-range dependence is still present beneath the noise in the observation process $(Y_n)_n$. Such dependencies imply that traditional statistical estimation techniques do not apply and may not work. As Lalley and Nobel state in [71], “though some features of denoising can be found in more traditional statistical problems such as errors in variables regression, deconvolution, and measurement error modeling (c.f. [12]), other features distinguish it from these problems and require new methods of analysis.” In particular, they cite the facts that the covariates X_n are deterministically related (as opposed to i.i.d. or mixing), the noise is often bounded (as opposed to Gaussian), and the noise distribution itself is often unknown.

EXAMPLE 4.1. Let $\mathcal{X} = [0, 1]$, and let $T_a : \mathcal{X} \rightarrow \mathcal{X}$ be given by $T_a(x) = ax(1-x)$, with a in $\mathcal{A} = [0, 4]$. This family of maps, known as the *logistic family*, has been extensively studied in a variety of settings. For $a \in [0, 1]$, it is known that for all x in $[0, 1]$, the iterates $T_a^n(x)$ tend to 0 as n tends to infinity. We say that a parameter value a has an attracting periodic orbit $\{p_0, \dots, p_{N-1}\}$ if $T_a(p_i) = p_{i+1}$ (with indices interpreted modulo N) and $|(T_a^N)'(p_i)| < 1$. For such parameter values, the iterates $T_a^n(x_0)$ of Lebesgue almost every initial point x_0 will tend to the periodic orbit $\{p_0, \dots, p_{N-1}\}$ as n tends to infinity. It is known [40, 74] that the set of parameter values that have an attracting periodic orbit is open and dense in $[0, 4]$. On the other hand, there are parameter values that give rise to very different asymptotic dynamics. In particular, we say that a

parameter value a has an absolutely continuous invariant measure (acim) μ_a if μ_a is absolutely continuous with respect to Lebesgue and μ_a is an invariant measure for T_a . In such cases, it can be shown that the iterates $T_a^n(x_0)$ of Lebesgue almost every initial point x_0 equidistribute with respect to μ_a . Intuitively, the presence of μ_a produces seemingly stochastic behavior, which is often referred to as chaos. Jakobson showed in [48] that the set of parameter values that have an acim has positive measure in $[0, 4]$, and Lyubich eventually showed in [75] that Lebesgue almost every parameter in $[0, 4]$ either has an attracting periodic orbit or an acim.

In most of the papers cited in this section, this family of maps is taken as a standard testing ground for parameter estimation methods. Generally, it is assumed that the observational noise is additive (*i.e.* $f_a(x, \epsilon) = x + \sigma(a)\epsilon$).

4.1 Noise reduction

One basic approach to parameter estimation in the observational noise case is to reduce the noise and then apply parameter estimation methods. If the noise can be uniformly and sufficiently reduced, then these approaches will be approximately as successful as the estimation method applied to the noiseless case. For example, the positive results in [69, 70, 71] might be combined with a parameter estimation method in order to produce consistent estimates. Among the results contained in these works, the main positive result of [71] is the most general, and we state it as follows.

A homeomorphism F of a compact metric space (Λ, d) is said to be expansive with separation threshold Δ if for every $x \neq y$ in Λ , there exists n in \mathbb{Z} such that $d(F^n(x), F^n(y)) > \Delta$. In the work [71], the authors consider an initial condition x and let $x_i = F^i(x)$. Also, they define a particular denoising algorithm which, given noisy additive noisy observations $(y_i)_{i=0}^{n-1}$, produces estimates $\hat{x}_{i,n}$ of the true states x_i . In this context, the main positive result may be stated as the following theorem.

THEOREM 4.2 ([71]). *Let $F : \Lambda \rightarrow \Lambda$ be an expansive homeomorphism with separation threshold $\Delta > 0$. Suppose that the noise process $(\epsilon_n)_n$ satisfies $|\epsilon_n| \leq \Delta/5$ for every n . If $k = k(n) \rightarrow \infty$ and $k/\log(n) \rightarrow 0$ as n tends to infinity, then*

$$\frac{1}{n-2k} \sum_{i=k}^{n-k} |\hat{x}_{i,n} - x_i| \rightarrow 0, \quad \text{as } n \rightarrow \infty$$

with probability 1 for every initial point x in Λ (with respect to any F invariant Borel probability measure).

By allowing a slight modification to their estimation scheme, the authors also show that under the same hypotheses

$$\max_{\log(n) \leq i \leq n - \log(n)} |\hat{x}_{i,n} - x_i| \rightarrow 0, \quad \text{as } n \rightarrow \infty$$

with probability 1 for almost every initial point x in Λ .

Of course, the task of removing the noise might itself be difficult or in some cases even impossible, as witnessed by the negative results in [69, 70, 71] and the related results in [52, 53, 55, 56]. Here we state the main negative result in [71]. A

pair of points x and x' is said to be strongly homoclinic for the homeomorphism F if

$$\sum_{k \in \mathbb{Z}} d(F^k(x), F^k(x')) < \infty.$$

THEOREM 4.3 ([71]). *Suppose the stationary distribution for the noise process $(\epsilon_n)_n$ is unbounded (or has sufficiently large support). If x and x' are strongly homoclinic, then for every measurable function $\phi : \prod_n \mathcal{X} \rightarrow \mathbb{R}^d$,*

$$\mathbb{E} \left[|\phi((y_n)_n) - x| - |\phi((y'_n)_n) - x'| \right] > 0.$$

In other words, even with access to the entire observation sequence, any state estimation or denoising scheme will fail with positive probability.

Lalley and Nobel also point out that despite the fact that in such cases the problem of asymptotic denoising is impossible, it might still be possible to obtain consistent parameter estimates. In fact, they state that such examples might provide an interesting avenue for further study (see Question 7.1).

In addition to the works mentioned so far in this section, the following works discuss the problem of denoising or smoothing data in the presence of only observational noise: [22, 39, 66, 68, 76, 77, 108].

4.2 Introduction to likelihoods and related methods

We begin with the work of Berliner [6, 7], sets the stage for most of the work that has followed. In these works, the author is mostly concerned with the observational noise setting (4.1)-(4.2). The likelihood function is given by

$$L(x_0, a) = p(y_0^{n-1} | x_0, a),$$

where $p(y_0^{n-1} | x_0, a)$ denotes the likelihood of observing y_0^{n-1} given the parameter choice a and the true initial condition x_0 (i.e. $p(\cdot | x_0, a)$ is the probability density for the observation process conditional on x_0 and a). The maximum likelihood (ML) method for estimating the parameter a amounts to defining the following maximum likelihood estimator (MLE):

$$(4.3) \quad \hat{a}_n = \operatorname{argmax}_a \max_{x_0} L(x_0, a).$$

It will be useful to find an explicit form for the likelihood function in the case that (I) the observational noise sequence $(\epsilon_n)_n$ is assumed to be i.i.d. normal with zero mean and unit variance, and (II) the observation function f_a takes the form $f_a(x, \epsilon) = x + \sigma(a)\epsilon$. The function $\sigma(a)$ allows one to set the variance of the noise according to the parameter a . In this case, we have

$$L(x_0, a) = \left(\sigma(a) \sqrt{2\pi} \right)^{-n} \exp \left(- \sum_{k=0}^{n-1} (y_k - T_a^k(x_0))^2 / (2\sigma^2(a)) \right)$$

and the corresponding log-likelihood function is given by

$$(4.4) \quad \log L(x_0, a) = -n \log \left(\sigma(a) \sqrt{2\pi} \right) - \sum_{k=0}^{n-1} (y_k - T_a^k(x_0))^2 / (2\sigma^2(a)).$$

A significant portion of the work on parameter estimation following Berliner has involved optimization of this log likelihood function, even when the noise is not necessarily Gaussian and thus its interpretation as a log likelihood function is no longer valid.

As discussed in [98], no existing statistical results apply to the ML method in this setting. With the above notation, the main difficulty in the current setting is that T_a^k is a non-stationary function of k . Standard statistical results on the performance of the ML method apply when the likelihood function has no such dependence on k (or is periodic with respect to k), but these results do not apply *a priori* in the current setting.

The Bayesian approach assumes a prior distribution (density) for x_0 and a , written as $\pi(x_0, a)$. Given the data y_0^{n-1} , the posterior distribution is then

$$\pi(x_0, a | y_0^{n-1}) = \frac{p(y_0^{n-1} | x_0, a) \pi(x_0, a)}{\int p(y_0^{n-1} | x, a) \pi(x, a) dx da}.$$

In these basic definitions, Berliner considers three main methods of parameter estimation: maximum likelihood estimation, minimization of a cost function (which is often chosen to be the negative of the log likelihood function) and Bayesian estimation. One of Berliner's main points is that when the system (4.1)-(4.2) is chaotic, the likelihood function will also typically be chaotic, in the sense that it will be extremely jagged. The rough nature of these likelihood functions makes all three of the above methods of statistical estimation computationally very expensive, and much of the work following Berliner has been motivated by the need to mitigate this difficulty. Beyond these computational difficulties, we would like to emphasize that to our knowledge there are no general results concerning the consistency of any of these likelihood-based methods.

4.3 Variations on likelihood based methods

A common method of parameter estimation in practice is to minimize some cost function C with respect to the parameters. Given the observations $(y_k)_{k=0}^{n-1}$, such methods employ the following estimators:

$$\hat{a}_n = \operatorname{argmin}_a \min_{x_0} C(x_0, a, (y_k)_{k=0}^{n-1}),$$

where $C(x_0, a, (y_k)_{k=0}^{n-1})$ somehow measures the discrepancy of the observations and the system trajectory having parameter a and initial state x_0 .

As we mentioned in the previous section, the most basic cost function is the least squares cost function

$$(4.5) \quad C_{LS}(x_0, a, (y_k)_{k=0}^{n-1}) = \sum_{k=0}^{n-1} (y_k - T_a^k(x_0))^2.$$

Perhaps due to the sensitive dependence of C_{LS} on x_0 and the additional computational expense incurred by minimizing C_{LS} over x_0 , several authors considered minimization of a one-step least squares cost function, given by

$$(4.6) \quad C_{OSLS}(a, (y_k)_{k=0}^{n-1}) = \sum_{k=0}^{n-2} (y_{k+1} - T_a(y_k))^2,$$

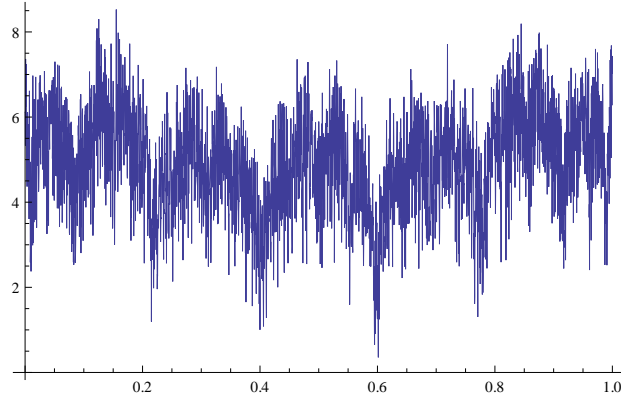


FIG 1. *Least Squares cost function for x_0 in logistic family as a function of $x \in [0, 1]$ given $n = 20$ observations, true initial value $x_0 = .4$ and true parameter $a = 4$.*

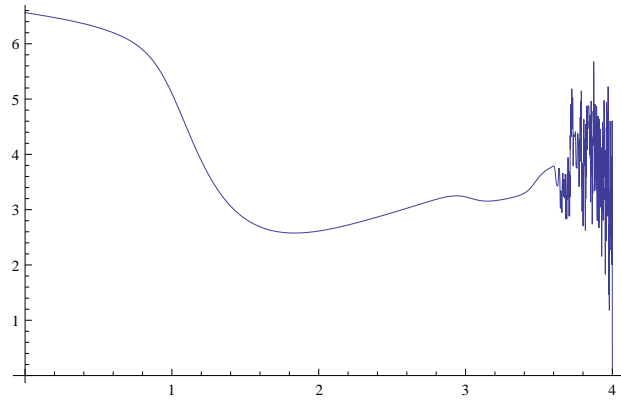


FIG 2. *Least Squares cost function for parameter a in logistic family as a function of $a \in [0, 4]$ given $n = 20$ observations, true initial value $x_0 = .4$ and true parameter $a = 4$.*

which does not depend on any initial condition x_0 . This cost function may appear to be the familiar least squares function from regression analysis, but as Kostelich [67] recognized, it suffers from the problem of errors in variables (*c.f.* [13, 38]). The problem of errors in variables is that the errors are not independent as is assumed by the cost function. Viewing C_{OSLS} from the perspective of traditional regression, we see that y_k appears to play the role of the independent variable and y_{k+1} plays the role of the dependent variable, but both y_k and y_{k+1} contain noise according to the model (4.1)-(4.2). It is well-known that the problem of errors in variables can lead to asymptotically biased results, and therefore we should not expect minimization of C_{OSLS} to give consistent estimates of the parameter a .

In response to the errors in variables problem, Jaeger and Kantz [47, 61] propose a “solution” of the problem, which amounts to minimizing the following cost function that has since gone by the name “total least squares” cost function:

$$(4.7) \quad C_{TLS}(a, (y_k)_{k=0}^{n-1}) = \sum_{k=0}^{n-1} \min_{y \in \mathcal{X}} |(y_k, y_{k+1}) - (y, T_a(y))|^2.$$

Note that this approach essentially ignores the dynamics altogether, and instead focuses on minimizing the sum of orthogonal distances between the graph of T_a

and the points (y_k, y_{k+1}) in $\mathcal{X} \times \mathcal{X}$. In order to include some aspect of the dynamics, they further modify their cost function to find local shadowing trajectories by considering cost functions of the form

$$(4.8) \quad C_{MTLS}(a) = \sum_{k=0}^{n-s-1} \min_y |(y_k, \dots, y_{k+s}) - (y, \dots, T_a^s(y))|.$$

Here s is a parameter of the method; it is the number of steps over which one considers the local shadowing trajectories. If one asks for global shadowing trajectories, corresponding to $s = n - 2$, then this modified total least squares cost function is equivalent to the original least squares cost function C_{LS} .

McSharry and Smith [81] consider the one step cost function C_{OSLS} given by (4.6). They prove that in the case of the logistic map with a specific parameter value, the minimization of this cost function produces biased estimates, even with infinitely many observations. Their proposed solution involves minimizing the cost function given by

$$(4.9) \quad C_{MS}(a) = - \sum_{k=0}^{n-1} \log \left(\int \exp \left(- \frac{d_k^2(x)}{2\epsilon^2} \right) \mu_a(dx) \right),$$

where $d_k^2(x) = |(y_k, y_{k+1}) - (x, T_a(x))|^2$, ϵ is the variance of the noise process $(\epsilon_n)_n$, and μ_a is a particular invariant measure for the map T_a . They argue that the minimum of C_{MS} provides more reliable parameter estimates due to its inclusion of information regarding the invariant measure μ_a . It is perhaps a shortcoming of this method that one must know the variance of the noise process and the invariant measure μ_a in order to calculate $C_{MS}(a)$. In practice, the authors suggest approximating the integral with respect to μ_a by a sum over a long piece of trajectory simulated from the model in the hopes that this approximation will be close to the integral by the ergodic theorem. The authors provide numerical evidence that C_{MS} provides better parameter estimates than either C_{OSLS} or C_{TLS} , although again no theoretical results are available to justify this comparison.

Meyer and Christensen [83], following up on the work of McSharry and Smith [81], propose to model the system using a combined noise state-space model of the form (2.1)-(2.2), and proceed via an MCMC algorithm for performing the inference. In particular, they take a Bayesian approach, modeling both the true states X_n and the parameters a as unknown variables. They assume that the process $(X_n)_n$ forms a Markov chain (by adding dynamical noise to the model). Then they compute posterior probabilities of the unobserved variables using the Gibbs sampler and the Metropolis-Hastings algorithm.

In his paper [51], titled “Chaotic-time-series reconstruction by the Bayesian paradigm: Right results by wrong methods,” Judd discusses the Bayesian approach of Meyer and Christensen [83], and argues that their approach might work, but for “accidental” reasons. In particular, he objects to the fact that Meyer and Christensen have replaced the deterministic model by a stochastic model, and he claims to formulate the “correct” Bayesian approach for the deterministic model, which he acknowledges is essentially that given by Berliner [6] (presented in Section 4.2). He argues that their model only appears to give correct results because it happens to find shadowing trajectories of the true system. (We remark that an ϵ shadowing trajectory for a sequence $(z_n)_n$ of states in \mathcal{X} is a true orbit $(x_n)_n$ of

the system such that $d(x_n, z_n) < \epsilon$ for all n .) In the end, he argues for methods based on a direct search for shadowing trajectories, which purportedly require significantly less computational effort than the Bayesian approach of Meyer and Christensen. Such methods are often referred to as gradient descent methods; for further reading about these methods, see [54, 104] and references therein.

Pisarenko and Sornette [98] consider the parameter estimation methods discussed above, as well as the method of moments. They point out that the method of moments seems to be the only method so far considered whose asymptotic consistency has been rigorously proved on even a single example. They also provide a careful analysis of the work of McSharry and Smith. This analysis sheds light on some errors, both quantitative and qualitative, in the work [81]. In order to provide a useful estimation procedure, they propose a “pure” likelihood method, in which they cut the time-series data into n_1 sub-intervals of length n_2 and perform ML estimation on each interval independently. In this method, the resulting n_1 parameter estimates are averaged to produce a single estimate. The motivation behind their method seems to be the following. A theoretically true/pure ML method involves treating x_0 as a parameter to be estimated (as in Section 4.2), but the chaotic nature of the system means that the system forgets its initial condition exponentially quickly, which implies that it cannot be reliably estimated. Hence, they arrive at the method of chopping the time series into smaller pieces (which hopefully still contain useful information about the initial condition of each piece) and using the pure ML method on each piece. The statistician interested in mathematical rigor is likely to find this work rewarding to read. A word of caution: they conclude their article by stating that “the situation is rather hopeless for the establishment of a meaningful statistical theory of estimation using the continuous theory of classical statistics to such discontinuous objects as the invariant measures of chaotic dynamical systems.”

Smirnov *et al* [112] note that the piecewise ML method of Pisarenko and Sornette [98] suffers from significant bias and potentially large variance, since it relies on chopping the data into many small subsets. In the case of one-dimensional maps, the authors propose a method based on backwards iteration of the map. They interpret their method as also relying on a ML principle, and they claim (with numerical support but no proof) that their method is asymptotically consistent with variance typically decreasing like n^{-2} , where n is the length of the observed time series.

The work of Horbelt and Timmer [43] seeks to quantify the rate of convergence of parameter estimates to the true parameter value in the observational noise case as the number of observations grows. In the introduction, the authors claim that the MLE in this setting is unbiased and efficient, for which they refer the reader to an earlier version of the book *Theory of point estimation* by Lehmann and Casella [72], although it seems clear that this statement is mistaken, since in some cases it can be shown to be asymptotically biased. Nonetheless, the authors find numerical evidence for various scaling laws of the variance of the MLE.

The work of Nakamura *et al* proposes yet another parameter estimation method in the observational noise setting [87]. Here the authors suggest an iterative method that alternates between estimating the system states and the system parameters. In each of the optimization steps, they use somewhat standard techniques. To estimate the parameters, they minimize the cost function C_{OSLS} in

(4.6) with respect to the parameter a . To estimate the states, they state that one could use any filtering method, such as the extended Kalman filter [121, 122] or gradient descent noise reduction (also known as gradient descent state estimation) [55, 66, 104]. The novelty of their approach lies in the fact that they iterate between these two estimation steps.

4.4 Method of moments

Here we mention a method of parameter estimation that has been shown to be consistent at least for the logistic family, discussed in Example 4.1. For the observational noise model, this method, discussed in [98], appears to be the only method that has been proved to be consistent for at least one non-trivial example.

We consider the model (4.1)-(4.2), where $\mathcal{X} = [-1, 1]$, $\mathcal{A} = [0, 2]$, and $T_a(x) = 1 - ax^2$, which is change of coordinates of the family in Example 4.1. Assume that the underlying trajectory process $(X_n)_n$ is ergodic, which is the case if one assumes that X_0 is drawn from an ergodic invariant measure μ_a for the map T_a . Alternatively, one may assume that a is chosen such that T_a has an acim μ_a (as discussed in Example 4.1) and X_0 is drawn from Lebesgue measure. Also assume that the observational noise is additive (*i.e.* $Y_n = X_n + \epsilon_n$) and $(\epsilon_n)_n$ is i.i.d. Gaussian with mean 0 and variance ϵ^2 . For a sequence $(z_k)_{k=0}^{n-1}$, let $A_n(z_k) = \frac{1}{n} \sum_{k=0}^{n-1} z_k$ and for any $f : \mathbb{R} \rightarrow \mathbb{R}$, let $\mathbb{E}_{\mu_a}(f) = \int f(x) d\mu_a(x)$. Then by the ergodic theorem

$$(4.10) \quad \lim_{n \rightarrow \infty} A_n(Y_k) = \mathbb{E}_{\mu_a}(x)$$

$$(4.11) \quad \lim_{n \rightarrow \infty} A_n(Y_k^2) = \mathbb{E}_{\mu_a}(x^2)$$

$$(4.12) \quad \lim_{n \rightarrow \infty} A_n(Y_k^3) = \mathbb{E}_{\mu_a}(x^3) + 3\epsilon^2 \mathbb{E}_{\mu_a}(x)$$

$$(4.13) \quad \lim_{n \rightarrow \infty} A_n(Y_k Y_{k+1}) = \mathbb{E}_{\mu_a}(x) - a \mathbb{E}_{\mu_a}(x^3).$$

Also, averaging the equation $x_{n+1} = 1 - ax_n^2$, we obtain that

$$(4.14) \quad \mathbb{E}_{\mu_a}(x) = 1 - a \mathbb{E}_{\mu_a}(x^2).$$

Combining Equations (4.10)-(4.14), we arrive at the following estimates for the unknown parameters a , $\mathbb{E}_{\mu_a}(x)$, $\mathbb{E}_{\mu_a}(x^2)$, $\mathbb{E}_{\mu_a}(x^3)$, and ϵ :

$$\begin{aligned} \hat{a}_n &= \frac{A_n(Y_k Y_{k+1}) + 2A_n(Y_k) + 3(A_n(Y_k))^2}{3A_n(Y_k)(A_n(Y_k))^2 - A_n(Y_k^3)} \\ \mathbb{E}_{\mu_a}(\hat{x})_n &= A_n(Y_k) \\ \mathbb{E}_{\mu_a}(\hat{x}^2)_n &= A_n(Y_k^2) - \hat{\epsilon}_n \\ \mathbb{E}_{\mu_a}(\hat{x}^3)_n &= \frac{1}{\hat{a}_n} (A_n(Y_k) - A_n(Y_k Y_{k+1})) \\ \hat{\epsilon}_n &= \frac{A_n(Y_k^3) - \mathbb{E}_{\mu_a}(\hat{x}^3)_n}{3A_n(Y_k)} \end{aligned}$$

These estimates are consistent by the ergodic theorem, but they might converge quite slowly, as there is no general rate of convergence in the ergodic theorem.

5. DYNAMICAL NOISE ONLY

If only dynamical noise is present in the model (2.1)-(2.2), then the system (2.1)-(2.2) reduces to the following situation:

$$(5.1) \quad Y_n = f_a(X_n)$$

$$(5.2) \quad X_{n+1} = T_a(X_n, \delta_n),$$

where $(\delta_n)_n$ is a noise process, $T : \mathcal{A} \times \mathcal{X} \times \mathcal{N} \rightarrow \mathcal{X}$ is a parametrized family of noisy maps, and $f : \mathcal{A} \times \mathcal{X} \times \mathcal{N} \rightarrow \mathcal{Y}$ is a parametrized family of observation functions. The dynamical noise model has been studied in the dynamical systems literature under the name “random dynamical systems” (see [65] and references therein). The process $(X_n)_n$ forms a discrete-time Markov chain on the continuous state space \mathcal{X} (see the book of Meyn and Tweedie [84] and references therein). In this case, some of the estimation methods from the statistical literature on time series and state space models may apply.

Without using this Markov structure, Adams and Nobel have studied the non-parametric reconstruction of such systems from direct observations (*i.e.* $Y_n = X_n$) [88, 89]. In particular, they used adaptive histogram methods to show results similar to those regarding non-parametric reconstructions of systems with no noise, as in Section 3.1. These methods do not work in the observational noise case precisely because in that setting they suffer from the problem of errors in variables, as discussed in Section 4.3.

A common setting for random dynamical systems is to assume that there is a map $T : \mathcal{X} \rightarrow \mathcal{X}$, where \mathcal{X} is a compact manifold and T is smooth, with a “natural” invariant probability measure μ . In common examples, T might be a (non-uniformly) hyperbolic map and μ might have the property that almost every initial condition with respect a volume measure on the manifold equidistributes with respect to μ . In such cases, one typically adds dynamical noise as follows. Let $\epsilon > 0$. For each x in \mathcal{X} , let $\mathbb{P}_\epsilon(x, \cdot)$ be the uniform measure on the ball of radius ϵ about the point $T(x)$. Then the Markov chain corresponding to this random dynamical system is determined by viewing \mathbb{P}_ϵ as the transition kernel for the chain. Under some conditions, the chain corresponding to \mathbb{P}_ϵ will have a unique stationary distribution, μ_ϵ . A well-known result (see [65]) states that under certain conditions, the measure μ_ϵ converges to μ weakly as ϵ tends to 0. To the best of our knowledge, no theoretical work on parameter estimation has been conducted for this particular setting, perhaps making it an area ripe for progress. On the other hand, this setting may be viewed as a particularly degenerate version of the general state-space setting, in which there is no observational noise, and therefore all methods described in Section 6 may also be applied here.

6. GENERAL STATE SPACE MODELS

In this section we consider the full system (2.1)-(2.2), where both dynamical noise and observational noise are present. Specific versions of such models have long been considered in the statistics literature, where they are known as state space models [29]. The literature on state space models in both applied and theoretical statistics is extensive and [42, 99] are two excellent texts covering applied modeling on this topic. The models can be summarized as the study of hidden Markov models (HMMs) in general state-spaces. (For an article discussing the

connections between ergodic theory and finite state HMMs, see [9].) Theoretical understanding of general HMMs has been a challenge and rigorous statements on consistency in parameter estimation have only appeared recently [27] (see Section 6.2). Most of the work in this area has been devoted to the problem of state estimation or filtering, and even at a computational level the problem of parameter estimation is still largely unsolved. In this section we survey some of the most studied approaches to filtering and discuss parameter estimation where there are results.

6.1 Kalman filter and some generalizations

The simplest such models assume that the dynamics are linear and the noise is additive Gaussian:

$$\begin{aligned} X_{n+1} &= AX_n + B\delta_{n+1} \\ Y_n &= CX_n + D\epsilon_n, \end{aligned}$$

where here A , B , C , and D are all matrices of the appropriate dimension and $(\delta_n)_n$ and $(\epsilon_n)_n$ are independent i.i.d. Gaussian processes. In this case, the optimal solution to the state estimation or denoising problem is given by the well-known Kalman filter [31, 59]. Generalizations of the ideas behind Kalman filtering to non-parametric models have been an extensive area of research in Bayesian and frequentist inference [29, 37, 36, 86].

Conceptually, the simplest generalization of the Kalman filter to nonlinear models involves linearizing the models at each time point and then using the Kalman filter. This method is often called the *extended Kalman filter* (EKF) [49, 5]. While the Kalman filter is optimal in the sense that it is the minimal-variance unbiased estimator, the general EKF is known to be biased. Furthermore, due to the linearization of the model, the propagation of the error covariance estimates may behave quite poorly if the non-linear terms in the model are significant.

The unscented Kalman filter (UKF) [57, 58] provides a deterministic sampling scheme that has been observed to outperform the EKF. The basic idea behind the UKF is that instead of approximating the model by linearization, one ought to use the exact model but approximate the posterior distributions by Gaussian distributions. The sampling scheme is designed to insure that the first two moments of the posterior distributions match the first two moments of the approximating distributions. It is believed that the UKF outperforms the EKF because it may be viewed as an unbiased second-order method, whereas the EKF is a biased first-order method. Of course, the UKF is believed to have shortcomings of its own; in particular, it assumes that the posterior distributions are Gaussian, which is certainly not the case in general. Also, the number of samples required for the UKF is at least the dimension of the state space, and in high-dimensional settings this fact makes the UKF computationally intractable. A wide variety of Monte Carlo (MC) methods have been proposed to overcome these issues.

Another generalization of the Kalman filter is known as the ensemble Kalman filter (EnKF) [32, 11, 33]. This method is a Monte Carlo method that is particularly popular in the weather prediction community. In fact, this method may be thought of as a type of particle filter (see Section 6.4.1).

6.2 MLE for HMMs

If one is willing to consider point estimates of unknown parameters in a setting where the likelihood function is known, then one can consider the maximum likelihood method (MLE) for parameter estimation. Let us now state the main result of the paper [27], which gives sufficient conditions for the consistency of MLE in this context. Let $(X_k, Y_k)_{k=1}^\infty$ be a hidden Markov model (HMM) of the form (2.1)-(2.2). Let a^* denote a fixed parameter value in \mathcal{A} . Assume that the HMM with parameter a^* has a unique stationary distribution, and let \mathbb{P}_{a^*} be the corresponding stationary HMM. Denote by $p^\nu(y_0^n, a)$ the likelihood of the observations Y_0^n with initial distribution $X_0 \sim \nu$ and parameter a . Consistency of the maximum likelihood estimator (MLE) may now be stated in the following form: if $a_n = \operatorname{argmax}_a p^\nu(y_0^n, a)$, then a_n converges \mathbb{P}_{a^*} -a.s. to a^* as n tends to infinity. The main result of [27] gives some general conditions under which the MLE is consistent in this sense. A precise statement of these general conditions is beyond the scope of this survey.

6.3 Bayesian inference

Recall the Bayesian formulation of state space estimation or filtering. Here one assumes that the model (2.1)-(2.2) gives rise to probability densities $\mu(x_0)$, $p(x|x')$, and $q(y|x)$, which define the initial distribution, transition kernel, and marginal distribution of the observation process, respectively. The densities are with respect to some fixed reference measures denoted dx and dy . In this framework, we are given access to finitely many observations y_0^{n-1} , and we would like to estimate the true trajectory x_0^{n-1} . Our assumptions define likelihood functions

$$p(x_0^{n-1}) = \mu(x_0) \prod_{k=0}^{n-2} p(x_{k+1}|x_k),$$

and

$$p(y_0^{n-1}|x_0^{n-1}) = \prod_{k=0}^{n-1} q(y_k|x_k).$$

Given the observations y_0^{n-1} , the posterior distribution for X_0^{n-1} is given by

$$p(x_0^{n-1}|y_0^{n-1}) = \frac{p(x_0^{n-1}, y_0^{n-1})}{p(y_0^{n-1})},$$

where

$$\begin{aligned} p(x_0^{n-1}, y_0^{n-1}) &= p(x_0^{n-1})p(y_0^{n-1}|x_0^{n-1}) \\ p(y_0^{n-1}) &= \int p(x_0^{n-1}, y_0^{n-1}) dx_0^{n-1}. \end{aligned}$$

There are a few instances when these distributions may be calculated analytically, such as when the system is linear and the noise is Gaussian or when $\{X_n\}_n$ is a finite state Markov chain. Outside of these cases, there is no analytical method for calculating the posterior distribution, and therefore one seeks a numerical approximation for this distribution. With the significant advances in computational power in recent years, there has been a remarkable amount of

research devoted to finding efficient computational approaches to approximating such posterior distributions. In the remainder of this section we briefly discuss the linear Gaussian case and a few of its generalizations to nonlinear or non-Gaussian situations. Section 6.4 discusses some of the more recent computational approaches to filtering.

An interesting work in the Bayesian context is [110] where the author studies posterior consistency for dependent data from an information theoretic point of view. The author establishes posterior consistency for misspecified models under the assumption of asymptotic equipartition property. For finite state space ergodic models, this is implied by the Shannon-McMillan-Breiman theorem. It could be interesting and useful to extend the ideas from [110] to prove posterior consistency in parameter estimation for more general dynamical systems.

6.4 Inference for dynamical systems via simulation based methods

In the general non-linear, non-Gaussian state-space setting of (2.1)-(2.2), the posterior distributions for x_0^{n-1} are not available in closed form, as they involve some integrals for which no analytical evaluation methods exist. In order to perform inference in this setting, a great deal of effort has been devoted to developing sophisticated computational algorithms for sampling from these posterior distributions. One general idea is to use Monte Carlo (MC) methods to estimate the integrals of interest. It is worth emphasizing that there has been a huge amount of work in this direction, and we do not claim to provide a comprehensive survey of all the relevant results. For an introduction to MC methods, see the book [105].

6.4.1 MCMC methods, SMC and Particle Filters. If one cannot sample from the posterior distribution directly, then one often turns to Markov chain Monte Carlo (MCMC) methods. For a discussion of such methods, see the books [105, 124] and references therein. Such methods have been used for parameter estimation in dynamical systems (*e.g.*, [21]).

Traditional Monte Carlo or MCMC methods may be used to perform “batch” inference, *i.e.* when all of the observations are available at once and one would like to estimate $p(x_0^{n-1}|y_0^{n-1})$ for fixed n , although even in this setting they might be prohibitively computationally expensive. When the goal is to perform “on-line” or sequential inference, or in an effort to try to reduce the computational expense, one might try sequential Monte Carlo methods (SMC) and their many variations. A particularly popular version of these methods is known as particle filtering. For a well-written, thorough introduction to the principles of sequential Monte Carlo (SMC) and particle filtering methods, see the recent tutorial by Doucet and Johansen [28]. For an incomplete list of works concerning SMC and particle filtering, as well as their adaptations to parameter estimation, see [11, 24, 26, 32, 33, 34, 44, 60, 73, 85, 90, 100, 115]. The basic idea is that the posterior distributions of interest are approximated by a finite collection of N samples, called particles, which are recursively propagated through the model. The main theoretical advantage of these methods is that one is often able to establish the convergence of the approximations to the true posterior distributions as the number of particles N tends to infinity.

6.4.2 ABC methods. Most of the methods mentioned previously in this section rely on explicit knowledge and evaluation of the likelihood function. In many situations, such as in high dimensional complex models, the likelihood

function may not be available or is computationally expensive to evaluate. In such scenarios, a simple computational method called approximate Bayesian computation (ABC) offers a powerful alternative to conduct statistical inference. ABC was first proposed as a philosophical argument in [107] and introduced to population genetics in [117]. Since then these methods have become extremely popular in many applied fields. A partial list of references include [23, 25, 80, 96, 101, 103, 111, 119, 125]. A good review with applications to filtering is [46]. Briefly speaking, in ABC methods one first draws a parameter value θ^* from the prior distribution and generates synthetic data from the likelihood model corresponding to θ^* . If the synthetic data “is similar to” the observed data (measured in some metric) up to a prespecified tolerance then θ^* is accepted as a draw from the (approximate) posterior distribution. Choosing the metric and the tolerance level are difficult problems, but partial results are known ([35]).

An important point to note is that in many examples, a summary statistic instead of the original data set is used for matching. This clearly results in loss of information (and sometimes even results in invalid inference; see [106]) and thus raises the interesting question about when one can perform consistent model selection using the ABC methodology. In [78] a sufficient criteria is worked out, but clearly more needs to be done especially in the context of dynamical systems.

7. OPEN QUESTIONS AND FUTURE DIRECTIONS

Here we list some open questions related to parameter inference in dynamical systems and discuss possible future research directions.

The first question examines if parameter estimation is possible even if denoising is impossible.

QUESTION 7.1. As shown by Lalley and Nobel [71], there are instances in which state estimation or denoising in the observational noise setting is impossible. Is it possible to exhibit a family of topological dynamical systems (X, T_a) on a compact metric space X such that consistent denoising is (provably) impossible but consistent parameter estimation is nonetheless (provably) possible?

The most common method in theory and practice for parameter estimation is ML. It is open if ML in the observational noise setting is consistent. A related question is can the approach to proving consistency results for HMMs in [27] be adapted to the observational noise setting.

QUESTION 7.2. Recall the definition of the ML estimator of the parameter a given in (4.3):

$$\hat{a}_n = \operatorname{argmax}_a \max_{x_0} L(x_0, a),$$

where $L(x_0, a)$ is the likelihood of x_0 and a conditional on the observations $(y_k)_{k=0}^{n-1}$. In the observational noise case setting (4.1)-(4.2), what are necessary and sufficient conditions on the system such that \hat{a}_n converges to a with probability 1 (for almost every initial condition x_0 with respect to an ergodic measure μ)? If necessary and sufficient conditions are out of reach given current tools, partial answers to this question in the form of general sufficient conditions might also be interesting.

In order to get finite sample error bounds, one would also like to know about the deviations of the MLE from its average. This line of reasoning leads to the following question.

QUESTION 7.3. In the observational noise setting (4.1)-(4.2), under which conditions on the system is it true that the MLE is asymptotically normal in the observational noise setting?

Since the method of moments presented in Section 4.4 is currently the only example to our knowledge for which consistency of any parameter estimation method can be proved in the observational noise setting, it is worth considering how it might be generalized.

QUESTION 7.4. Can the method of moments presented in Section 4.4 for the logistic family be generalized? Under what conditions is it applicable and consistent?

In the combined noise setting of Section 6, it is still the case that the issue of parameter inference has not been satisfactorily resolved. Certainly any filtering method may be trivially extended to a parameter estimation algorithm by extending the state space to include the parameters, but in such cases the degeneracy of the extended system typically causing the filtering methods to fail. Let us paraphrase a question in [28].

QUESTION 7.5. Under what conditions on the model are there efficient algorithms for parameter estimation in the general state space setting? What theoretical guarantees can be given to justify such algorithms?

The range of applications of statistical inference methods for deterministic dynamical systems seems to be increasing rapidly. These systems present significant new challenges, since the deterministic systems may have very long-range dependency structures. It would be a significant breakthrough if methods could be developed that provided asymptotically consistent algorithms for parameter estimation; moreover, one would like to have finite-size sample bounds on the accuracy of these algorithms. Given the difficulty of dealing with the long-range dependencies present in general in the observational noise model, it appears likely that the traditional methods of parameter inference may not work particularly well in this setting, and therefore new ideas and methods should be developed.

One possible approach would be to consider a weakened notion of consistency. For example, one could consider a parameter estimation method to be consistent if it returns a set of plausible parameters that asymptotically contains the true parameter. Such weakened notions of consistency might be necessary for providing some theoretical justification of parameter estimation algorithms when achieving strong consistency appears out of reach.

Let us close with one recent development in the field of dynamical systems and ergodic theory that might be useful in obtaining such rates of convergence. The concentration inequalities mentioned at the end of Section 3.3 provide a powerful method for obtaining finite sample error bounds for a wide class of statistical estimators for a wide class of dynamical systems. One might hope that these

concentration inequalities can be used to get rigorous error bounds for parameter estimation algorithms.

Acknowledgments The authors would like to thank Andrew Nobel, Ramon van Handel, John Harer, Konstantin Mischaikow, Christian Robert, Mark Girolami and Andrew Stuart for discussions, comments and help with references. SM and KM would like to acknowledge AFOSR FA9550-10-1-0436 and NSF DMS-1045153 for partial support. SM would also like to acknowledge NSF CCF-1049290 for partial support. NSP would like to thank NSF for partial support through the grant NSF DMS-1107070.

REFERENCES

- [1] Henry D. I. Abarbanel. *Analysis of observed chaotic data*. Institute for Nonlinear Science. Springer-Verlag, New York, 1996.
- [2] Henry D. I. Abarbanel, Daniel R. Creveling, Reza Farsian, and Mark Kostuk. Dynamical state and parameter estimation. *SIAM J. Appl. Dyn. Syst.*, 8(4):1341–1381, 2009.
- [3] Henry D. I. Abarbanel, Daniel R. Creveling, and James M. Jeanne. Estimation of parameters in nonlinear systems using balanced synchronization. *Phys. Rev. E* (3), 77(1):016208, 14, 2008.
- [4] Terrence M. Adams and Andrew B. Nobel. Finitary reconstruction of a measure preserving transformation. *Israel J. Math.*, 126:309–326, 2001.
- [5] B.D.O. Anderson and J.B. Moore. *Optimal Filtering*. Englewood Cliffs, 1979.
- [6] L. Mark Berliner. Likelihood and Bayesian prediction of chaotic systems. *J. Amer. Statist. Assoc.*, 86(416):938–952, 1991.
- [7] L. Mark Berliner. Statistics, probability and chaos. *Statist. Sci.*, 7(1):69–122, 1992. With discussion and a rejoinder by the author.
- [8] Boris P. Bezruchko and Dmitry A. Smirnov. *Extracting knowledge from time series*. Springer Series in Synergetics. Springer, Heidelberg, 2010. An introduction to nonlinear empirical modeling.
- [9] Mike Boyle and Karl Petersen. Hidden Markov processes in the context of symbolic dynamics. In *Entropy of hidden Markov processes and connections to dynamical systems*, volume 385 of *London Math. Soc. Lecture Note Ser.*, pages 5–71. Cambridge Univ. Press, Cambridge, 2011.
- [10] Michael Brin and Garrett Stuck. *Introduction to dynamical systems*. Cambridge University Press, Cambridge, 2002.
- [11] G. Burgers, P. Jan van Leeuwen, and G. Evensen. Analysis scheme in the ensemble kalman filter. *Mon. Wea. Rev.*, 126:1719–1724, 1998.
- [12] R. J. Carroll, D. Ruppert, and L. A. Stefanski. *Measurement error in nonlinear models*, volume 63 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 1995.
- [13] Raymond J. Carroll, David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. *Measurement error in nonlinear models*, volume 105 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, Boca Raton, FL, second edition, 2006. A modern perspective.
- [14] Kung-Sik Chan and Howell Tong. *Chaos: a statistical perspective*. Springer Series in Statistics. Springer-Verlag, New York, 2001.
- [15] Sangit Chatterjee and Mustafa R. Yilmaz. Chaos, fractals and statistics. *Statist. Sci.*, 7(1):49–68, 1992.
- [16] J.-R. Chazottes. Fluctuations of observables in dynamical systems: from limit theorems to concentration inequalities. arXiv:1201.3833v1, 2012.
- [17] J.-R. Chazottes, P. Collet, F. Redig, and E. Verbitskiy. A concentration inequality for interval maps with an indifferent fixed point. *Ergodic Theory Dynam. Systems*, 29(4):1097–1117, 2009.

- [18] J.-R. Chazottes, P. Collet, and B. Schmitt. Devroye inequality for a class of non-uniformly hyperbolic dynamical systems. *Nonlinearity*, 18(5):2323–2340, 2005.
- [19] J.-R. Chazottes, P. Collet, and B. Schmitt. Statistical consequences of the Devroye inequality for processes. Applications to a class of non-uniformly hyperbolic dynamical systems. *Nonlinearity*, 18(5):2341–2364, 2005.
- [20] J.-R. Chazottes and S. Gouezel. Optimal concentration inequalities for dynamical systems. arXiv:1111.0849v1, 2011.
- [21] Nelson Christensen, Renate Meyer, Lloyd Knox, and Ben Luey. Bayesian methods for cosmological parameter estimation from cosmic microwave background measurements. *Classical and Quantum Gravity*, 18(14):2677, 2001.
- [22] Mike Davies. Noise reduction schemes for chaotic time series. *Phys. D*, 79(2–4):174–192, 1994.
- [23] Thomas A. Dean, Sumeetpal S. Singh, Ajay Jasra, and Gareth w. Peters. Parameter estimation for hidden Markov models with intractable likelihoods. arXiv:1103.5399v1, 2011.
- [24] Pierre Del Moral. *Feynman-Kac formulae*. Probability and its Applications (New York). Springer-Verlag, New York, 2004. Genealogical and interacting particle systems with applications.
- [25] Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. An adaptive sequential monte carlo method for approximate bayesian computation. *Statistics and Computing*, pages 1–12, 2011.
- [26] Pierre Del Moral, Arnaud Doucet, and Sumeetpal Singh. Forward smoothing using Sequential Monte Carlo. arXiv:1012.5390v1, 2010.
- [27] Randal Douc, Eric Moulines, Jimmy Olsson, and Ramon van Handel. Consistency of the maximum likelihood estimator for general hidden Markov models. *Ann. Statist.*, 39(1):474–513, 2011.
- [28] Arnaud Doucet and Adam Johansen. *A tutorial on particle filtering and smoothing: fifteen years later*, chapter 8.2. Oxford University Press, 2011.
- [29] J. Durbin and S. J. Koopman. *Time series analysis by state space methods*, volume 24 of *Oxford Statistical Science Series*. Oxford University Press, Oxford, 2001.
- [30] J.-P. Eckmann and D. Ruelle. Ergodic theory of chaos and strange attractors. *Rev. Modern Phys.*, 57(3, part 1):617–656, 1985.
- [31] R. L. Eubank. *A Kalman filter primer*, volume 186 of *Statistics: Textbooks and Monographs*. Chapman & Hall/CRC, Boca Raton, FL, 2006.
- [32] Geir Evensen. Sequential data assimilation with a nonlinear quasi-geostrophic model using monte carlo methods to forecast error statistics. *J. Geophys. Res.*, 99:10143–10162, 1994.
- [33] Geir Evensen. The ensemble kalman filter: theoretical formulation and practical implementation. *Ocean Dynamics*, 53:343–367, 2003.
- [34] Paul Fearnhead. Markov chain monte carlo, sufficient statistics, and particle filters. *Journal of Computational and Graphical Statistics*, 11(4):pp. 848–862, 2002.
- [35] Paul Fearnhead and Dennis Prangle. Constructing summary statistics for approximate bayesian computation: semi-automatic approximate bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):419–474, 2012.
- [36] E. Fox, E.B. Sudderth, M.I. Jordan, and A.S. Willsky. Bayesian nonparametric inference of switching dynamic linear models. *Signal Processing, IEEE Transactions on*, 59(4):1569–1585, april 2011.
- [37] E.B. Fox, E.B. Sudderth, M.I. Jordan, and A.S. Willsky. Bayesian nonparametric methods for learning markov switching processes. *Signal Processing Magazine, IEEE*, 27(6):43–54, nov. 2010.
- [38] Wayne A. Fuller. *Measurement error models*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2006. Reprint of the 1987 original, Wiley-Interscience Paperback Series.
- [39] Jianbo Gao, H. Sultan, Jing Hu, and Wen-Wen Tung. Denoising nonlinear time series by adaptive filtering and wavelet shrinkage: A comparison. *Signal Processing Letters, IEEE*, 17(3):237–240, 2010.

- [40] Jacek Graczyk and Grzegorz Świątek. Generic hyperbolicity in the logistic family. *Ann. of Math. (2)*, 146(1):1–52, 1997.
- [41] Yonatan Gutman and Michael Hochman. On processes which cannot be distinguished by finite observation. *Israel J. Math.*, 164:265–284, 2008.
- [42] James D. Hamilton. *Time-series analysis*. Princeton University Press, 1 edition, January 1994.
- [43] W. Horbelt and J. Timmer. Asymptotic scaling laws for precision of parameter estimates in dynamical systems. *Phys. Lett. A*, 310(4):269–280, 2003.
- [44] Edward L. Ionides, Anindya Bhadra, Yves Atchadé, and Aaron King. Iterated filtering. *Ann. Statist.*, 39(3):1776–1802, 2011.
- [45] Valerie Isham. Statistical aspects of chaos: a review. In *Networks and chaos—statistical and probabilistic aspects*, volume 50 of *Monogr. Statist. Appl. Probab.*, pages 124–200. 1993.
- [46] Marin J.-M., Pudlo P., Robert C.P., and R. Ryder. Approximate bayesian computational methods. *Statistics and Computing*, 2(21):289–291.
- [47] L. Jaeger and H. Kantz. Unbiased reconstruction of the dynamics underlying a noisy chaotic time series. *Chaos*, 6:440–450, 1996.
- [48] M. V. Jakobson. Absolutely continuous invariant measures for one-parameter families of one-dimensional maps. *Comm. Math. Phys.*, 81(1):39–88, 1981.
- [49] A.H. Jazwinski. *Stochastic Processes and Filtering Theory*. Academic Press, 1970.
- [50] Jens Ledet Jensen. Chaotic dynamical systems with a view towards statistics: a review. In *Networks and chaos—statistical and probabilistic aspects*, volume 50 of *Monogr. Statist. Appl. Probab.*, pages 201–250. 1993.
- [51] Kevin Judd. Chaotic-time-series reconstruction by the bayesian paradigm: Right results by wrong methods. *Phys. Rev. E*, 67:026212, Feb 2003.
- [52] Kevin Judd. Nonlinear state estimation, indistinguishable states, and the extended Kalman filter. *Phys. D*, 183(3-4):273–281, 2003.
- [53] Kevin Judd. Failure of maximum likelihood methods for chaotic dynamical systems. *Phys. Rev. E*, 75:036210, Mar 2007.
- [54] Kevin Judd. Shadowing pseudo-orbits and gradient descent noise reduction. *Journal of Nonlinear Science*, 18:57–74, 2008.
- [55] Kevin Judd and Leonard Smith. Indistinguishable states. I. Perfect model scenario. *Phys. D*, 151(2-4):125–141, 2001.
- [56] Kevin Judd and Leonard A. Smith. Indistinguishable states ii: The imperfect model scenario. *Physica D: Nonlinear Phenomena*, 196(3-4):224 – 242, 2004.
- [57] S.J. Julier and J.K. Uhlmann. A general method for approximating nonlinear transformations of probability distributions. Technical report, Department of Engineering Science, Oxford University, 1996.
- [58] S.J. Julier and J.K. Uhlmann. A new extension of the kalman filter to nonlinear systems. In *Proc. of AeroSense: The 11th International Symposium on Aerospace/Defense Sensing, Simulation, and Controls*, 1997.
- [59] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME—Journal of Basic Engineering*, 82(Series D):35–45, 1960.
- [60] N. Kantas, A. Doucet, S.S. Singh, and J.M. Maciejowski. An overview of sequential monte carlo methods for parameter estimation in general state-space models. In *Proceedings of the IFAC System Identification Meeting*, 2009.
- [61] Holger Kantz and Lars Jaeger. Improved cost functions for modelling of noisy chaotic time series. *Physica D: Nonlinear Phenomena*, 109(12):59–69, 1997.
- [62] Holger Kantz and Thomas Schreiber. *Nonlinear time series analysis*. Cambridge University Press, Cambridge, second edition, 2004.
- [63] Anatole Katok and Boris Hasselblatt. *Introduction to the modern theory of dynamical systems*, volume 54 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 1995. With a supplementary chapter by Katok and Leonardo Mendoza.
- [64] Yuri Kifer. *Random perturbations of dynamical systems*, volume 16 of *Progress in Probability and Statistics*. Birkhäuser Boston Inc., Boston, MA, 1988.

- [65] Yuri Kifer and Pei-Dong Liu. Random dynamics. In *Handbook of dynamical systems. Vol. 1B*, pages 379–499. Elsevier B. V., Amsterdam, 2006.
- [66] E. Kostelich and T. Schreiber. Noise reduction schemes for chaotic time-series data: a survey of common methods. *Phys. Rev. E*, 48:1752–1763, 1993.
- [67] Eric J. Kostelich. Problems in estimating dynamics from data. *Physica D: Nonlinear Phenomena*, 58(1-4):138–152, 1992.
- [68] Eric J. Kostelich and James A. Yorke. Noise reduction: finding the simplest dynamical system consistent with the data. *Phys. D*, 41(2):183–196, 1990.
- [69] Steven P. Lalley. Beneath the noise, chaos. *Ann. Statist.*, 27(2):461–479, 1999.
- [70] Steven P. Lalley. Removing the noise from chaos plus noise. In *Nonlinear dynamics and statistics (Cambridge, 1998)*, pages 233–244. Birkhäuser Boston, Boston, MA, 2001.
- [71] Steven P. Lalley and A. B. Nobel. Denoising deterministic time series. *Dyn. Partial Differ. Equ.*, 3(4):259–279, 2006.
- [72] E.L. Lehmann and G. Casella. *Theory of point estimation*. Springer, New York, 1998.
- [73] Hedibert F. Lopes and Ruey S. Tsay. Particle filters and bayesian inference in financial econometrics. *Journal of Forecasting*, 30(1):168–209, 2011.
- [74] Mikhail Lyubich. Combinatorics, geometry and attractors of quasi-quadratic maps. *Ann. of Math. (2)*, 140(2):347–404, 1994.
- [75] Mikhail Lyubich. Almost every real quadratic map is either regular or stochastic. *Ann. of Math. (2)*, 156(1):1–78, 2002.
- [76] G. Manjunath, S. Sivaji Ganesh, and G. V. Anand. Topology-based denoising of chaos. *Dyn. Syst.*, 24(4):501–516, 2009.
- [77] G. Manjunath, S. Sivaji Ganesh, and G. V. Anand. Denoising signals corrupted by chaotic noise. *Commun. Nonlinear Sci. Numer. Simul.*, 15(12):3988–3997, 2010.
- [78] J.-M. Marin, N.S. Pillai, C. P. Robert, and J. Rousseau. Relevant statistics for Bayesian model choice. *ArXiv e-prints*, October 2011.
- [79] Anil Maybhatte and R. E. Amritkar. Use of synchronization and adaptive control in parameter estimation from a time series. *Phys. Rev. E*, 59:284–293, Jan 1999.
- [80] Trevelyan McKinley, Alex R. Cook, and Robert Deardon. Inference in epidemic models without likelihoods. *The International Journal of Biostatistics*, 5(1):24, 2009.
- [81] Patrick E. McSharry and Leonard A. Smith. Better nonlinear models from noisy data: Attractors with maximum likelihood. *Phys. Rev. Lett.*, 83:4285–4288, Nov 1999.
- [82] Alistair I. Mees, editor. *Nonlinear dynamics and statistics*. Birkhäuser Boston Inc., Boston, MA, 2001. Selected papers from the workshop held at Cambridge University, Cambridge, September 1998.
- [83] Renate Meyer and Nelson Christensen. Bayesian reconstruction of chaotic dynamical systems. *Physical Review E*, 62, 2000.
- [84] Sean Meyn and Richard L. Tweedie. *Markov chains and stochastic stability*. Cambridge University Press, Cambridge, second edition, 2009. With a prologue by Peter W. Glynn.
- [85] C. Mukherjee and M. West. Sequential monte carlo in model comparison: Example in cellular dynamics in systems biology. In *JSM Proceedings, Section on Bayesian Statistical Science. Alexandria, VA: American Statistical Association*, pages 1274–1287, 2009.
- [86] Chiranjit Mukherjee. *Bayesian Modelling and Computation in Dynamic and Spatial Systems*. PhD thesis, Duke University, Durham, North Carolina, 2011.
- [87] Tomomichi Nakamura, Yoshito Hirata, Kevin Judd, Devin Kilminster, and Michael Small. Improved parameter estimation from noisy time series for nonlinear dynamical systems. *Internat. J. Bifur. Chaos Appl. Sci. Engrg.*, 17(5):1741–1752, 2007.
- [88] Andrew Nobel. Consistent estimation of a dynamical map. In *Nonlinear dynamics and statistics (Cambridge, 1998)*, pages 267–280. Birkhäuser Boston, Boston, MA, 2001.
- [89] Andrew B. Nobel and Terrence M. Adams. Estimating a function from ergodic samples with additive noise. *IEEE Trans. Inform. Theory*, 47(7):2895–2902, 2001.
- [90] Jimmy Olsson, Olivier Cappé, Randal Douc, and Eric Moulines. Sequential Monte Carlo smoothing with application to parameter estimation in nonlinear state space models. *Bernoulli*, 14(1):155–179, 2008.
- [91] D. S. Ornstein and B. Weiss. Statistical properties of chaotic systems. *Bull. Amer. Math. Soc. (N.S.)*, 24(1):11–116, 1991. With an appendix by David Fried.

- [92] Donald Ornstein and Benjamin Weiss. Entropy is the only finitely observable invariant. *J. Mod. Dyn.*, 1(1):93–105, 2007.
- [93] Donald S. Ornstein and Benjamin Weiss. How sampling reveals a process. *Ann. Probab.*, 18(3):905–930, 1990.
- [94] N.H. Packard, J.P. Crutchfield, J.D. Farmer, and R.S. Shaw. Geometry from a time series. *Phys. Rev. Lett.*, 45(9):712–715, 1980.
- [95] U. Parlitz. Estimating model parameters from time series by autosynchronization. *Phys. Rev. Lett.*, 76:1232–1235, Feb 1996.
- [96] Gareth W. Peters, Mario V. Wüthrich, and Pavel V. Shevchenko. Chain ladder method: Bayesian bootstrap versus classical bootstrap. *Insurance: Mathematics and Economics*, 47(1):36 – 51, 2010.
- [97] Karl Petersen. *Ergodic theory*, volume 2 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1989. Corrected reprint of the 1983 original.
- [98] V. F. Pisarenko and D. Sornette. Statistical methods of parameter estimation for deterministically chaotic time series. *Phys. Rev. E*, 69:036122, Mar 2004.
- [99] A. Pole, M. West, and P. J. Harrison. *Applied Bayesian Forecasting & Time Series Analysis*. Chapman-Hall, 1994.
- [100] G. Poyiadjis, A. Doucet, and S.S. Singh. Particle approximations of the score and observed information matrix in state space models with application to parameter estimation. 98(1):65–80, 2011.
- [101] J K Pritchard, M T Seielstad, A Perez-Lezaun, and M W Feldman. Population growth of human y chromosomes: a study of y chromosome microsatellites. *Molecular Biology and Evolution*, 16(12):1791–1798, 1999.
- [102] John C. Quinn, Paul H. Bryant, Daniel R. Creveling, Sallee R. Klein, and Henry D. I. Abarbanel. Parameter and state estimation of experimental chaotic systems using synchronization. *Phys. Rev. E (3)*, 80(1):016201, 17, 2009.
- [103] Oliver Ratmann, Christophe Andrieu, Carsten Wiuf, and Sylvia Richardson. Model criticism based on likelihood-free inference, with an application to protein network evolution. *Proceedings of the National Academy of Sciences*, 106(26):10576–10581, 2009.
- [104] David Ridout and Kevin Judd. Convergence properties of gradient descent noise reduction. *Physica D: Nonlinear Phenomena*, 165(1-2):26 – 47, 2002.
- [105] Christian P. Robert and George Casella. *Monte Carlo statistical methods*. Springer Texts in Statistics. Springer-Verlag, New York, second edition, 2004.
- [106] Christian P. Robert, J.M. Cornuet, J.M. Marin, and N.S. Pillai. Lack of confidence in approximate bayesian computational (abc) model choice. *PNAS*, 108(37):15112–15117, 2011.
- [107] D Rubin. Bayesianly justiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, 12(4):1151–1172, 1984.
- [108] Tim Sauer. A noise reduction method for signals from nonlinear systems. *Phys. D*, 58(1-4):193–201, 1992. Interpretation of time series from nonlinear systems (Warwick, 1991).
- [109] Tim Sauer, James A. Yorke, and Martin Casdagli. Embedology. *J. Statist. Phys.*, 65(3-4):579–616, 1991.
- [110] Cosma R. Shalizi. Dynamics of bayesian updating with dependent data and misspecified models. *Electron. J. Stat.*, (3):10391074, 2009.
- [111] S. A. Sisson, Y. Fan, and Mark M. Tanaka. Sequential monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 104(6):1760–1765, 2007.
- [112] Dmitry A. Smirnov, Vladislav S. Vlaskin, and Vladimir I. Ponomarenko. Estimation of parameters in one-dimensional maps from noisy chaotic time series. *Phys. Lett. A*, 336(6):448–458, 2005.
- [113] J. Stark, D. S. Broomhead, M. E. Davies, and J. Huke. Delay embeddings for forced systems. II. Stochastic forcing. *J. Nonlinear Sci.*, 13(6):519–577, 2003.
- [114] Thomas Stemler and Kevin Judd. A guide to using shadowing filters for forecasting and state estimation. *Phys. D*, 238(14):1260–1273, 2009.
- [115] G. Storvik. Particle filters for state-space models with the presence of unknown static parameters. *Signal Processing, IEEE Transactions on*, 50(2):281–289, feb 2002.

- [116] Floris Takens. Detecting strange attractors in turbulence. In *Dynamical systems and turbulence, Warwick 1980 (Coventry, 1979/1980)*, volume 898 of *Lecture Notes in Math.*, pages 366–381. Springer, Berlin, 1981.
- [117] S. Tavaré, D. J. Balding, R. C. Griffiths, and P. Donnelly. Inferring coalescence times from dna sequence data. *Genetics*, 145(2):505–518, 1997.
- [118] Howell Tong. *Nonlinear time series*, volume 6 of *Oxford Statistical Science Series*. The Clarendon Press Oxford University Press, New York, 1990. A dynamical system approach, With an appendix by K. S. Chan, Oxford Science Publications.
- [119] T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M. Stumpf. Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Interface*, 6(31):187–202, 2009.
- [120] Henning U. Voss, Jens Timmer, and Jürgen Kurths. Nonlinear dynamical system identification from uncertain and indirect measurements. *Internat. J. Bifur. Chaos Appl. Sci. Engrg.*, 14(6):1905–1933, 2004.
- [121] D. Walker and A. Mees. Noise reduction of chaotic systems by kalman filtering and by shadowing. *Int. J. Bifurcation and Chaos*, 7(3):769–779, 1997.
- [122] D. Walker and A. Mees. Reconstructing nonlinear dynamics by extended kalman filtering. *Int. J. Bifurcation and Chaos*, 8(3):557–569, 1998.
- [123] Peter Walters. *An introduction to ergodic theory*, volume 79 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1982.
- [124] M. West and P. J. Harrison. *Bayesian Forecasting and Dynamic Models*. Springer Verlag, 2nd edition, 1997.
- [125] Richard D. Wilkinson. Approximate bayesian computation (abc) gives exact results under the assumption of model error. arXiv:0811.3355v1, 2008.
- [126] Lai-Sang Young. Statistical properties of dynamical systems with some hyperbolicity. *Ann. of Math. (2)*, 147(3):585–650, 1998.
- [127] Lai-Sang Young. What are SRB measures, and which dynamical systems have them? *J. Statist. Phys.*, 108(5-6):733–754, 2002. Dedicated to David Ruelle and Yasha Sinai on the occasion of their 65th birthdays.
- [128] Wenwu Yu, Guanrong Chen, Jinde Cao, Jinhu Lü, and Ulrich Parlitz. Parameter identification of dynamical systems from time series. *Phys. Rev. E*, 75:067201, Jun 2007.